On power law distributions in large-scale taxonomies

Rohit Babbar, Cornelia Metzig, Ioannis Partalas, Eric Gaussier, Massih-Reza Amini

On Power Law Distributions in Large-scale Taxonomies

Rohit Babbar, Cornelia Metzig, Ioannis Partalas,
Eric Gaussier and Massih-Reza Amini
Université Grenoble Alpes, CNRS
F-38000 Grenoble, France
First.Last@imag.fr

Abstract

In many of the large-scale physical and social complex systems phenomena fat-tailed distributions occur, for which different generating mechanisms have been proposed. In this paper, we study models of generating power law distributions in the evolution of large-scale taxonomies such as Open Directory Project, which consist of websites assigned to one of tens of thousands of categories. The categories in such taxonomies are arranged in tree or DAG structured configurations having parent-child relations among them. We first quantitatively analyse the formation process of such taxonomies, which leads to power law distribution as the stationary distributions. In the context of designing classifiers for large-scale taxonomies, which automatically assign unseen documents to leaf-level categories, we highlight how the fat-tailed nature of these distributions can be leveraged to analytically study the space complexity of such classifiers. Empirical evaluation of the space complexity on publicly available datasets demonstrates the applicability of our approach.

1 Introduction

With the tremendous growth of data on the web from various sources such as social networks, online business services and news networks, structuring the data into conceptual taxonomies leads to better scalability, interpretability and visualization. Yahoo! directory, the open directory project (ODP) and Wikipedia are prominent examples of such web-scale taxonomies. The Medical Subject Heading hierarchy of the National Library of Medicine is another instance of a large-scale taxonomy in the domain of life sciences. These taxonomies consist of classes arranged in a hierarchical structure with parent-child relations among them and can be in the form of a rooted tree or a directed acyclic graph. ODP for instance, which is in the form of a rooted tree, lists over 5 million websites distributed among close to 1 million categories and is maintained by close to 100,000 human editors. Wikipedia, on the other hand, represents a

more complicated directed graph taxonomy structure consisting of over a million categories. In this context, large-scale hierarchical classification deals with the task of automatically assigning labels to unseen documents from a set of target classes which are represented by the leaf level nodes in the hierarchy.

In this work, we study the distribution of data and the hierarchy tree in large-scale taxonomies with the goal of modelling the process of their evolution. This is undertaken by a quantitative study of the evolution of large-scale taxonomy using models of preferential attachment, based on the famous model proposed by Yule [33] and showing that throughout the growth process, the taxonomy exhibits a fat-tailed distribution. We apply this reasoning to both category sizes and tree connectivity in a simple joint model. Formally, a random variable X is defined to follow a power law distribution if for some positive constant a, the complementary cumulative distribution is given as follows:

$$P(X > x) \propto x^{-a}$$

Power law distributions, or more generally fat-tailed distributions that decay slower than Gaussians, are found in a wide variety of physical and social complex systems, ranging from city population, distribution of wealth to citations of scientific articles [23]. It is also found in network connectivity, where the internet and Wikipedia are prominent examples [27, 7]. Our analysis in the context of large-scale web-taxonomies leads to a better understanding of such large-scale data, and also leveraged in order to present a concrete analysis of space complexity for hierarchical classification schemes. Due to the ever increasing scale of training data size in terms of the number of documents, feature set size and number of target classes, the space complexity of the trained classifiers plays a crucial role in the applicability of classification systems in many applications of practical importance.

The space complexity analysis presented in this paper provides an analytical comparison of the trained model for hierarchical and flat classification, which can be used to select the appropriate model a-priori for the classification problem at hand, without actually having to train any models. Exploiting the power law nature of taxonomies to study the training time complexity for hierarchical Support Vector Machines has been performed in [32, 19]. The authors therein justify the power law assumption only *empirically*, unlike our analysis in Section 3 wherein we describe the generative process of large-scale web taxonomies more concretely, in the context of similar processes studied in other models. Despite the important insights of [32, 19], space complexity has not been treated formally so far.

The remainder of this paper is as follows. Related work on reporting power law distributions and on large scale hierarchical classification is presented in Section 2. In Section 3, we recall important growth models and quantitatively justify the formation of power laws as they are found in hierarchical large-scale web taxonomies by studying the evolution dynamics that generate them. More specifically, we present a process that jointly models the growth in the size of categories, as well as the growth of the hierarchical tree structure. We derive

from this growth model why the class size distribution at a given level of the hierarchy also exhibits power law decay. Building on this, we then appeal to Heaps' law in Section 4, to explain the distribution of features among categories which is then exploited in Section 5 for analysing the space complexity for hierarchical classification schemes. The analysis is empirically validated on publicly available DMOZ datasets from the Large Scale Hierarchical Text Classification Challenge (LSHTC)[1] and patent data (IPC) [2] from World Intellectual Property Organization. Finally, Section 6 concludes this work.

2 Related Work

Power law distributions are reported in a wide variety of physical and social complex systems [22], such as in internet topologies. For instance [11, 7] showed that internet topologies exhibit power laws with respect to the in-degree of the nodes. Also the size distribution of website categories, measured in terms of number of websites, exhibits a fat-tailed distribution, as empirically demonstrated in [32, 19] for the Open Directory Project (ODP). Various models have been proposed for the generation power law distributions, a phenomenon that may be seen as fundamental in complex systems as the normal distribution in statistics [25]. However, in contrast to the straight-forward derivation of normal distribution via the central limit theorem, models explaining power law formation all rely on an approximation. Some explanations are based on multiplicative noise or on the renormalization group formalism [28, 30, 16]. For the growth process of large-scale taxonomies, models based on preferential attachment are most appropriate, which are used in this paper. These models are based on the seminal model by Yule [33], originally formulated for the taxonomy of biological species, detailed in section 3. It applies to systems where elements of the system are grouped into classes, and the system grows both in the number of classes, and in the total number of elements (which are here documents or websites). In its original form, Yule's model serves as explanation for power law formation in any taxonomy, irrespective of an eventual hierarchy among categories. Similar dynamics have been applied to explain scaling in the connectivity of a network, which grows in terms of nodes and edges via preferential attachment [2]. Recent further generalizations apply the same growth process to trees [17, 14, 29]. In this paper, describe the approximate power-law in the child-to-parent category relations by the model by Klemm et al. [17]. Furthermore, we combine this formation process in a simple manner with the original Yule model in order to explain also a power law in category sizes, i.e. we provide a comprehensive explanation for the formation process of large-scale web taxonomies such as DMOZ. From the second, we infer a third scaling distribution for the number of features per category. This is done via the empirical Heaps's law [10], which describes the scaling relationship between text length and the size of its vocabulary.

Some of the earlier works on exploiting hierarchy among target classes for

[1] http://lshtc.iit.demokritos.gr/
[2] http://web2.wipo.int/ipcpub/

the purpose of text classification have been studied in [18, 6] and [8] wherein the number of target classes were limited to a few hundreds. However, the work by [19] is among the pioneering studies in hierarchical classification towards addressing web-scale directories such as Yahoo! directory consisting of over 100,000 target classes. The authors analyse the performance with respect to accuracy and training time complexity for flat and hierarchical classification. More recently, other techniques for large-scale hierarchical text classification have been proposed. Prevention of error propagation by applying *Refined Experts* trained on a validation set was proposed in [4]. In this approach, bottom-up information propagation is performed by utilizing the output of the lower level classifiers in order to improve classification at top level. The deep classification method proposed in [31] first applies hierarchy pruning to identify a much smaller subset of target classes. Prediction of a test instance is then performed by re-training Naive Bayes classifier on the subset of target classes identified from the first step. More recently, Bayesian modelling of large-scale hierarchical classification has been proposed in [15] in which hierarchical dependencies between the parent-child nodes are modelled by centring the prior of the child node at the parameter values of its parent.

In addition to prediction accuracy, other metrics of performance such as prediction and training speed as well as space complexity of the model have become increasingly important. This is especially true in the context of challenges posed by problems in the space of Big Data, wherein an optimal trade-off among such metrics is desired. The significance of prediction speed in such scenarios has been highlighted in recent studies such as [3, 13, 24, 5]. The prediction speed is directly related to space complexity of the trained model, as it may not be possible to load a large trained model in the main memory due to sheer size. Despite its direct impact on prediction speed, no earlier work has focused on space complexity of hierarchical classifiers.

Additionally, while the existence of power law distributions has been used for analysis purposes in [32, 19] no thorough justification is given on the existence of such phenomenon. Our analysis in Section 3, attempts to address this issue in a quantitative manner. Finally, power law semantics have been used for model selection and evaluation of large-scale hierarchical classification systems [1]. Unlike problems studied in classical machine learning sense which deal with a limited number of target classes, this application forms a blue-print on extracting hidden information in big data.

3 Power Law in Large-Scale Web Taxonomies

We begin by introducing the complementary cumulative size distribution for category sizes. Let N_i denote the size of category i (in terms of number of documents), then the probability that $N_i > N$ is given by

$$P(N_i > N) \propto N^{-\beta} \tag{1}$$

where $\beta > 0$ denotes the exponent of the power law distribution.[3] Empirically, it can be assessed by plotting the rank of a category's size against its size (see Figure 1) The derivative of this distribution, the category size probability density $p(N_i)$, then also follows a power law with exponent $(\beta+1)$, i.e. $p(N_i) \propto N_i^{-(\beta+1)}$.

Two of our empirical findings are a power law for both the complementary cumulative category size distribution and the counter-cumulative in-degree distribution, shown in Figures 1 and 2, for LSHTC2-DMOZ dataset which is a subset of ODP. The dataset[4] contains 394,000 websites and 27,785 categories. The number of categories at each level of the hierarchy is shown in Figure 3.

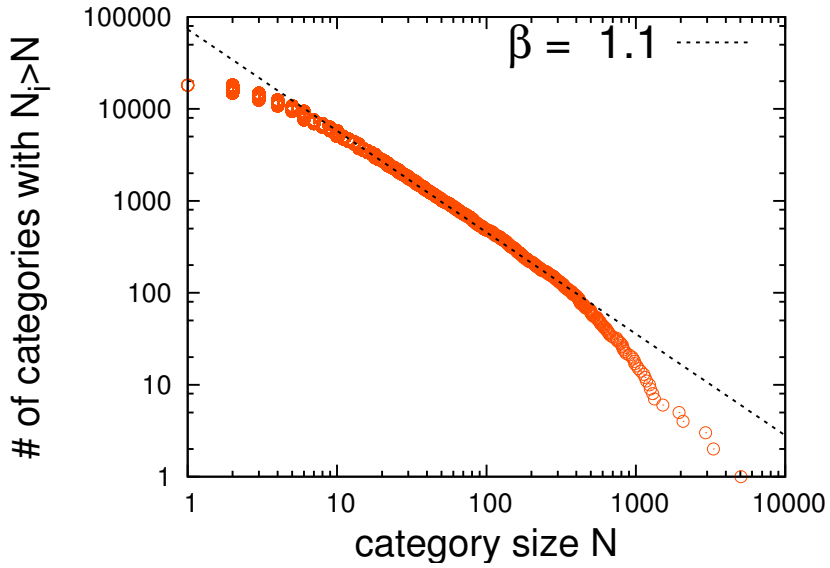

Figure 1: Category size vs rank distribution for the LSHTC2-DMOZ dataset.

We explain the formation of these two laws via models by Yule [33] and a related model by Klemm [17], detailed in sections 3.1 and 3.2, which are then related in section 3.3.

3.1 Yule's model

Yule's model describes a system that grows in two quantities, in elements and in classes in which the elements are assigned. It assumes that for a system having κ classes, the probability that a new element will be assigned to a certain class

[3] To avoid confusion, we denote the power law exponents for in-degree distribution and feature size distribution γ and δ.

[4] http://lshtc.iit.demokritos.gr/LSHTC2_datasets

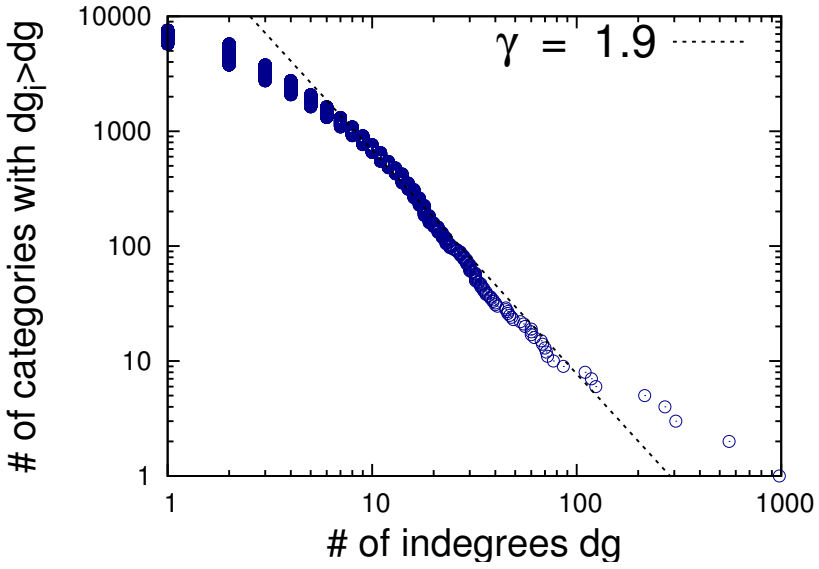

Figure 2: Indegree vs rank distribution for the LSHTC2-DMOZ dataset.

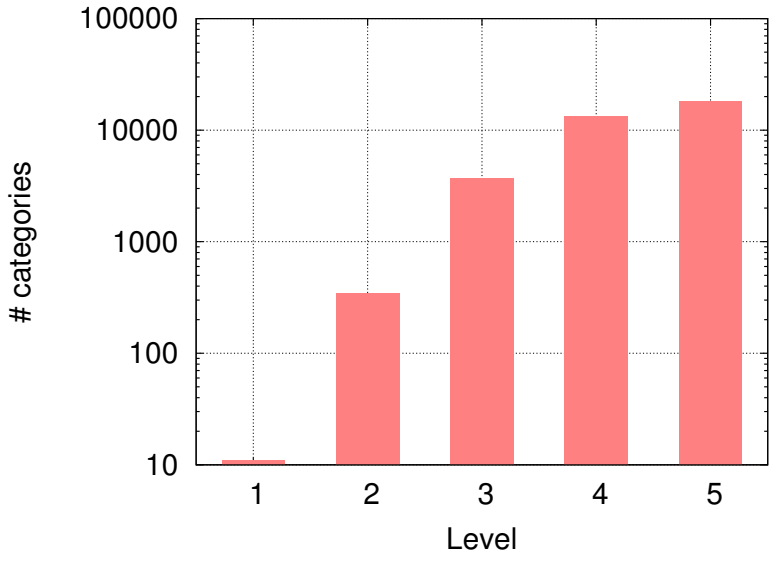

Figure 3: Number of categories at each level in the hierarchy of the LSHTC2-DMOZ database.

is proportional to its current size,

$$p(i) = \frac{N_i}{\sum_{i'=1}^{\kappa} N_{i'}} \tag{2}$$

Variables	
N_i	Number of elements in class i
dg_i	Number of subclasses of class i
d_i	Number of features of class i
κ	Total number of classes
DG	Total number of in-degrees (=subcategories)
$p_{N,\kappa}$	Fraction of classes having N elements when the total number of classes is κ
Constants	
m	Number of elements added to the system after which a new class is added
w	$\in [0,1]$ Probability that attachment of subcategories is preferential
Indices	
i	Index for the class

Table 1: Summary of notation used in Section 3

It further assumes that for every m elements that are added to the pre-existing classes in the system, a new class of size 1 is created[5].

The described system is constantly growing in terms of elements and classes, so strictly speaking, a stationary state does not exist [20]. However, a stationary distribution, the so-called Yule distribution, has been derived using the approach of the master equation with similar approximations by [26, 23, 17]. Here, we follow Newman [23], who considers as one time-step the duration between creation of two consecutive classes. From this follows that the average number of elements per class is always $m+1$, and the system contains $\kappa(m+1)$ elements at a moment where the number of classes is κ. Let $p_{N,\kappa}$ denote the fraction of classes having N elements when the total number of classes is κ. Between two successive time instances, the probability for a given pre-existing class i of size N_i to gain a new element is $mN_i/(\kappa(m+1))$. Since there are $\kappa\, p_{N,\kappa}$ classes of size N, the expected number such classes which gain a new element (and grow to size $(N+1)$) is given by :

$$\frac{mN}{\kappa(m+1)}\,\kappa\, p_{N,\kappa} = \frac{m}{(m+1)}\,N\, p_{N,\kappa} \tag{3}$$

The number of classes with N websites are thus fewer by the above quantity, but some which had $(N-1)$ websites prior to the addition of a new class have now one more website. This step depicting the change of the state of the system

[5]The initial size may be generalized to other small sizes; for instance Tessone et al. consider entrant classes with size drawn from a truncated power law [29] .

from κ classes to $(\kappa+1)$ classes is shown in Figure 4. Therefore, the expected number of classes with N documents when the number of classes is $(\kappa+1)$ is given by the following equation:

$$(\kappa+1)p_{N,(\kappa+1)} = \kappa\, p_{N,\kappa} + \frac{m}{m+1}[(N-1)(p_{(N-1),\kappa}) - Np_{N,\kappa}] \tag{4}$$

The first term in the right hand side of Equation 4 corresponds to classes with N documents when the number of classes is κ. The second term corresponds to the contribution from classes of size $(N-1)$ which have grown to size N, this is shown by the left arrow (pointing rightwards) in Figure 4. The last term corresponds to the decrease resulting from classes which have gained an element and have become of size $(N+1)$, this is shown by the right arrow (pointing rightwards) in Figure 4. The equation for the class of size 1 is given by:

$$(\kappa+1)p_{1,(\kappa+1)} = \kappa\, p_{1,\kappa} + 1 - \frac{m}{m+1}p_{1,\kappa} \tag{5}$$

As the number κ of classes (and therefore the number of elements $\kappa(m+1)$) in the system increases, the probability that a new element is classified into a class of size N, given by Equation 3, is assumed to remain constant and independent of κ. Under this hypothesis, the stationary distribution for class sizes can be determined by solving Equation 4 and using Equation 5 as the initial condition. This is given by

$$p_N = (1+1/m)B(N, 2+1/m) \tag{6}$$

where $B(.,.)$ is the beta distribution. Equation 6 has been termed *Yule distribution* [26]. Written for a continuous variable N, it has a power law tail:

$$p(N) \propto N^{-2-\frac{1}{m}}$$

From the above equation the exponent of the density function is between 2 and 3. Its cumulative size distribution $P(N_k > N)$, as given by Equation 1, has an exponent given by

$$\beta = (1+(1/m)) \tag{7}$$

which is between 1 and 2. The higher the frequency $1/m$ at which new classes are introduced, the bigger β becomes, and the lower the average class size. This exponent is stable over time although the taxonomy is constantly growing.

3.2 Preferential attachment models for networks and trees

A similar model has been formulated for network growth by Barabási and Albert [2], which explains the formation of a power law distribution in connectivity degree of nodes. It assumes that the networks grow in terms of nodes and edges, and that every newly added node to the system connects with a fixed number of edges to existing nodes. Attachment is again preferential, i.e. the

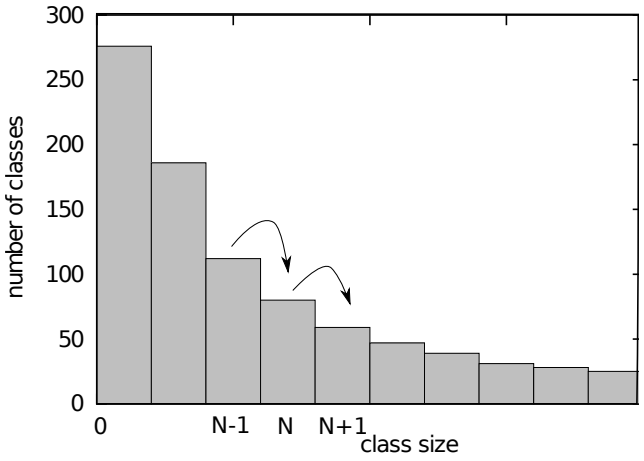

Figure 4: Illustration of Equation 4. Individual classes grow constantly i.e., move to the right over time, as indicated by arrows. A stationary distribution means that the height of each bar remains constant.

probability for a newly added node i to connect to a certain existing node j is proportional to its number of existing edges of node j.

A node in the Barabási-Albert (BA) model corresponds a class in Yule's model, and a new edge to two newly assigned element. Every added edge counts both to the degree of an existing node j, as well as to the newly added node i. For this reason the existing nodes j and the newly added node i grow always by the same number of edges, implying $m = 1$ and consequently $\beta = 2$ in the BA-model, *independently* of the number of edges that each new node creates.

The seminal BA-model has been extended in many ways. For hierarchical taxonomies, we use a preferential attachment model for trees by [17]. The authors considered growth via directed edges, and explain power law formation in the *in-degree*, i.e. the edges directed from children to parent in a tree structure. In contrast to the BA-model, newly added nodes and existing nodes do not increase their in-degree by the same amount, since new nodes start with an in-degree of 0. Leaf nodes thus cannot attract attachment of nodes, and preferential attachment alone cannot lead to a power-law. A small random term ensures that some nodes attach to existing ones independently of their degree, which is the analogous to the start of a new class in the Yule model. The probability v that a new node attaches as a child to the existing node i of with indegree dg_i becomes

$$v(i) = w\frac{d_i - 1}{DG} + (1 - w)\frac{1}{DG}\,, \tag{8}$$

where DG is the size of the system measured in the total number of in-degrees. $w \in [0, 1]$ denotes the probability that the attachment is preferential, $(1 - w)$ the probability that it is random to any node, independently of their numbers of

indegrees. As it has been done for the Yule process [26, 23, 14, 29], the stationary distribution is again derived via the master Equation 4. The exponent of the asymptotic power law in the in-degree distribution is $\beta = 1 + 1/w$.This model is suitable to explain scaling properties of the tree or network structure of large-scale web taxonomies, which have also been analysed empirically, for instance for subcategories of Wikipedia [7]. It has also been applied to directory trees in [14].

3.3 Model for hierarchical web taxonomies

We now apply these models to large-scale web taxonomies like DMOZ. Empirically, we uncovered two scaling laws: (a) one for the size distribution of leaf categories and (b) one for the indegree (child-to-parent link) distribution of categories (shown in Figure 2). These two scaling laws are linked in a non-trivial manner: a category may be very small or even not contain any websites, but nevertheless be highly connected. Since on the other hand (a) and (b) arise jointly, we propose here a model generating the two scaling laws in a simple generic manner. We suggest a combination of the two processes detailed in subsections 3.1 and 3.2 to describe the growth process: websites are continuously added to the system, and classified into categories by human referees. At the same time, the categories are not a mere set, but form a tree structure, which grows itself in two quantities: in the number nodes (categories) and in the number of in-degrees of nodes (child-to-parent links, i.e. subcategory-to-category links). Based on the rules for voluntary referees of the DMOZ how to classify websites, we propose a simple combined description of the process. Altogether, the database grows in *three* quantities:

(i) *Growth in websites.* New websites are assigned into categories i, with probability $p(i) \propto N_i$ (Figure 5). This assignment happens independently of the hierarchy level of category. However, only leaf categories may receive documents.

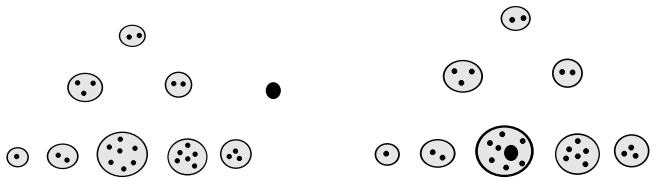

Figure 5: (i): A website is assigned to existing categories with $p(i) \propto N_i$.

(ii) *Growth in categories.* With probability $1/m$, the referees assign a website into a newly created category, at any level of the hierarchy (Figure 6).

This assumption would suffice to create a power law in the category size distribution, but since a tree-structure among categories exists, we also assume that the event of category creation is also attaching at particular

places to the tree structure. The probability $v(i)$ that a category is created as the child of a certain parent category i can depend in addition on the *in-degree* d_i of that category (see Equation 9).

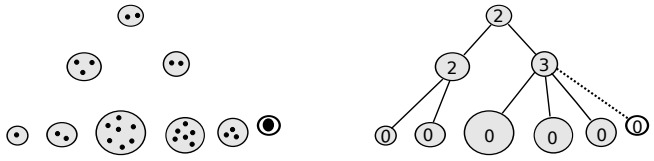

Figure 6: (ii): Growth in categories is equivalent to growth of the tree structure in terms of in-degrees.

(iii) *Growth in children categories.* Finally, the hierarchy may also grow in terms of levels, since with a certain probability $(1 - w)$, new children categories are assigned independently of the number of children, i.e. its in-degree d_i of the category i. (Figure 7). Like in [17], the attachment probability to a parent i is

$$v(i) = w\frac{dg_i - 1}{DG} + (1 - w)\frac{\epsilon_i}{DG}\,. \tag{9}$$

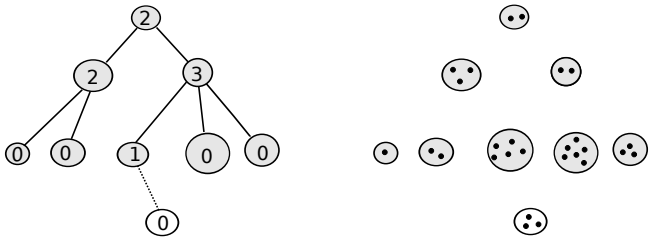

Figure 7: (iii): Growth in children categories.

Equation 8, where $\epsilon_i = 1$, would suffice to explain power law in-degrees dg_i and in category sizes N_i.

To link the two processes more plausibly, it can be assumed that the second term in Equation 9 denoting assignment of new 'first children' depends on the size N_i of parent categories,

$$\epsilon_i = \frac{N_i}{N}\,, \tag{10}$$

since this is closer to the rules by which the referees create new categories, but is not essential for the explanation of the power laws. It reflects that

the bigger a leaf category, the higher the probability that referees create a child category when assigning a new website to it.

To summarize, the central idea of this joint model is to consider two measures for the size of a category: the number of its websites N_i (which governs the preferential attachment of new websites), and its in-degree, i.e. the number of its children dg_i, which governs the preferential attachment of new categories. To explain the power law in the category sizes, assumptions (i) and (ii) are the requirements. For the power law in the number of indegrees, assumptions (ii) and (iii) are the requirements. The empirically found exponents $\beta = 1.1$ and $\gamma = 1.9$ yield a frequency of new categories 1/m=0.1 and a frequency of new indegrees $(1 - w) = 0.9$.

3.4 Other interpretations

Instead of assuming in Equations 9 and 10 that referees decide to open a single child category, it is more realistic to assume that an existing category is *restructured*, i.e. one or several child categories are created, and websites are moved into these new categories such that the parent category contains less websites or even none at all. If one of the new children categories inherits all websites of the parent category (see Figure 8), the Yule model applies directly. If the websites are partitioned differently, the model contains effective shrinking of categories. This is not described by the Yule model, and the master Equation 4 considers only growing categories. However, it has been shown [29, 21] that models including shrinking categories also lead to the formation of power laws. Further generalizations compatible with power law formation are that new categories do not necessarily start with one document, and that the frequency of new categories does not need to be constant.

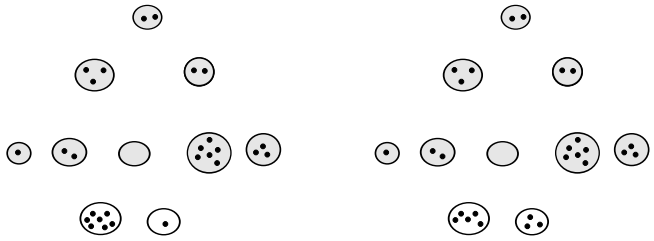

Figure 8: Model without and with shrinking categories. In the left figure, a child category inherits all the elements of its parent and takes its place in the size distribution.

3.5 Limitations

However, Figures 1 and 2 do not exhibit perfect power law decay for several reasons. Firstly, the dataset is limited. Secondly, the hypothesis that the assignment probability (Equation 2) depends uniquely on the size of a category

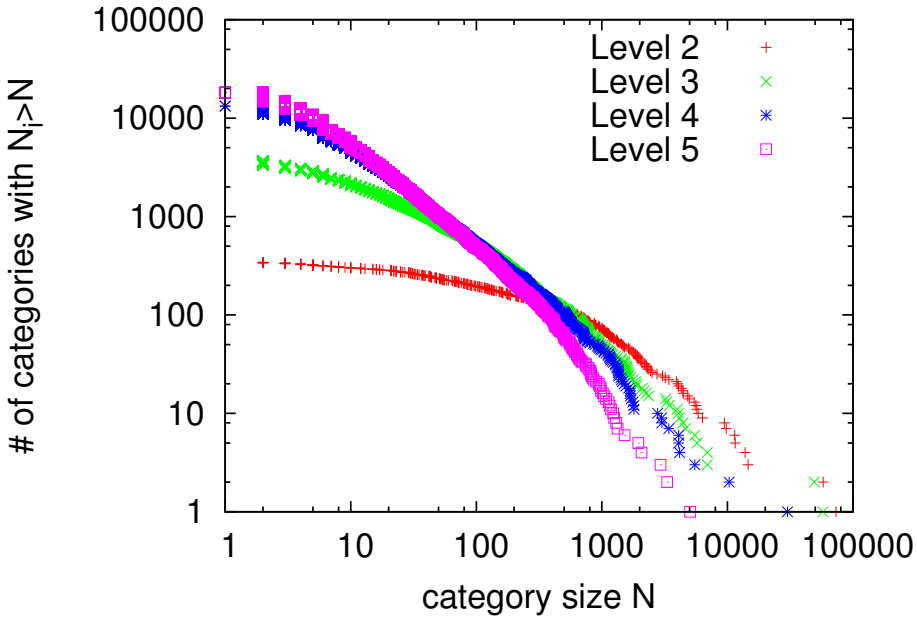

Figure 9: Category size distribution for each level of the LSHTC2-DMOZ dataset.

might be too strong for web directories, neglecting the change in importance of topics. In reality, big categories can exist which receive only few new documents or none at all. Dorogovtsev and Mendes [9] have studied this problem by introducing an assignment probability that decays exponentially with age. For a low decay parameter they show that the stronger this decay, the steeper the power law; for strong decay, no power law forms. A last reason might be that referees re-structure categories in ways strongly deviating from the rules (i) - (iii).

3.6 Statistics per hierarchy level

The tree-structure of a database allows also to study the sizes of class belonging to a given level of the hierarchy. As shown in Figure 3 the DMOZ database contains 5 levels of different size. If only classes on a given level l of the hierarchy are considered, we equally found a power law in category size distribution as shown in Figure 9. Per-level power law decay has also been found for the in-degree distribution. This result may equally be explained by the model introduced above: Equations 2 and 9 respectively, are valid also if instead of $p(k)$ one considers the conditional probability $p(l)p(i|l)$, where $p(l) = \frac{\sum_{i'=1,l}^{\kappa} N_{i',l}}{\sum_{i'=1}^{\kappa} N_{i'}}$ is the probability of assignment to a given level, and $p(i|l) = \frac{N_{i,l}}{\sum_{i'=1,l}^{\kappa} N_{i',l}}$ the probability of being assigned to a given class within that level. The formation

process may be seen as a Yule process *within* a level if $\sum_{i'=1,l}^{\kappa} N_{i',l}$ is used for the normalization in Equation 2, and this formation happens with probability $p(l)$ that a website gets assigned into level l. Thereby, the rate at m_l at which new classes are created need not be the same for every level, and therefore the exponent of the power law fit may vary from level to level. Power law decay for the per-level class size distribution is a straightforward corollary of the described formation process, and will be used in Section 5 to analyse the space complexity of hierarchical classifiers.

4 Relation between category size and number of features

Having explained the formation of two scaling laws in the database, a third one has been found for the number of features d_i in each category, $G(d)$ (see Figures 11 and 12). This is a consequence of both the category size distribution, shown (in Figure 1) in combination with another power law, termed Heaps' law [10]. This empirical law states that the number of distinct words R in a document is related to the length n of a document as follows

$$R(n) = Kn^{\alpha}\,, \tag{11}$$

where the empirical α is typically between 0.4 and 0.6. For the LSHTC2-DMOZ dataset, Figure 10 shows that for the collection of words and the collection of websites, similar exponents are found. An interpretation of this result is that the total number words in a category can be measured approximately by the number of websites in a category, although not all websites have the same length.

Figure 10 shows that bigger categories contain also more features, but this increase is weaker than the increase in websites. This implies that less very 'feature-rich' categories exist, which is also reflected in the high decay exponent $\delta = 1.9$ of a power-law fit in Figure 11, (compared to the slower decay of the category size distribution shown in figure 1 where $\beta = 1.1$). Catenation of the size distribution measured in features and Heaps' law yields again size distribution measured in websites: $P(i) = R(G(d_i))$, i.e. multiplication of the exponents yields that $\delta \cdot \alpha = 1.1$ which confirms our empirically found value $\beta = 1.1$.

5 Space Complexity of Large-Scale Hierarchical Classification

Fat-tailed distributions in large-scale web taxonomies highlight the underlying structure and semantics which are useful to visualize important properties of the data especially in big data scenarios. In this section we focus on the applications in the context of large-scale hierarchical classification, wherein the fit of power law distribution to such taxonomies can be leveraged to concretely analyse the

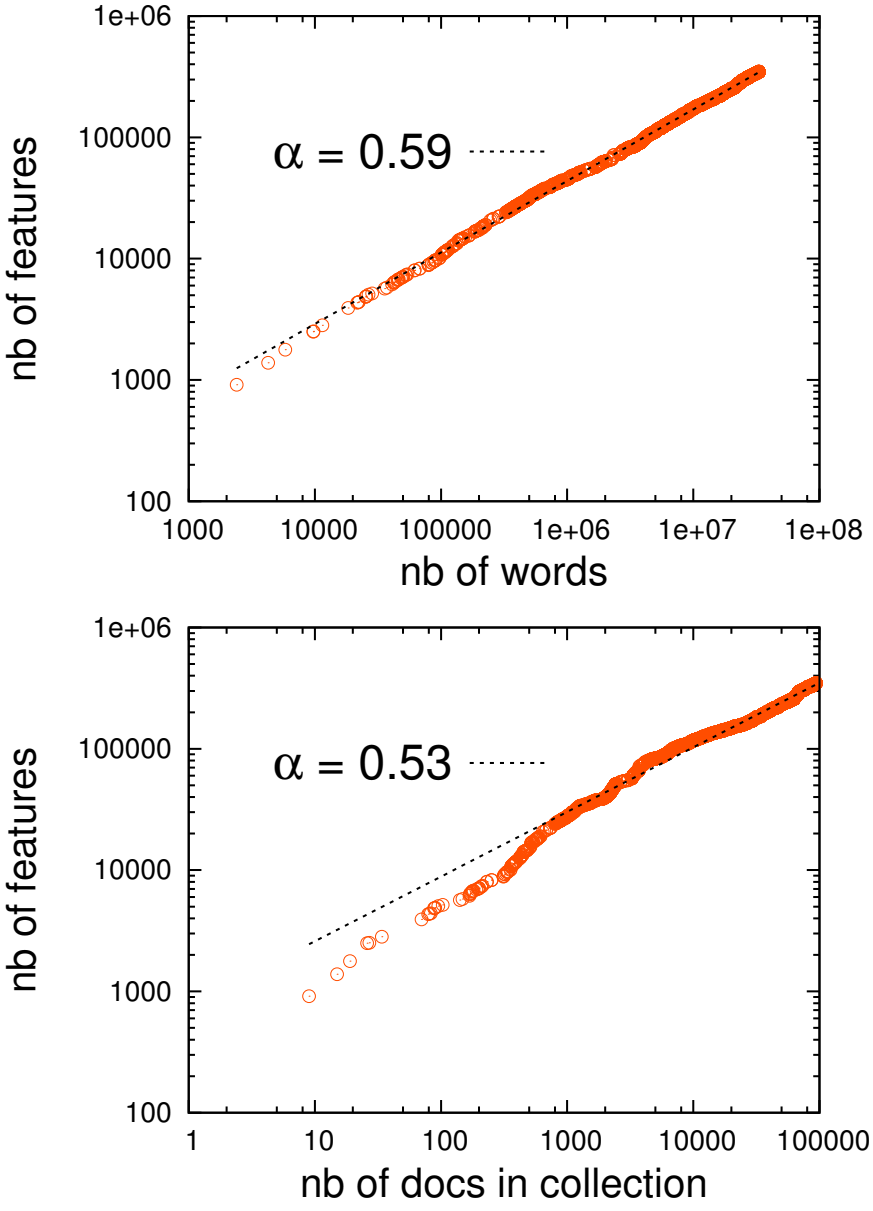

Figure 10: Heaps' law: number of distinct words vs. number of words, and vs number of documents.

space complexity of large-scale hierarchical classifiers in the context of a generic linear classifier deployed in top-down hierarchical cascade.

In the following sections we first present formally the task of hierarchical

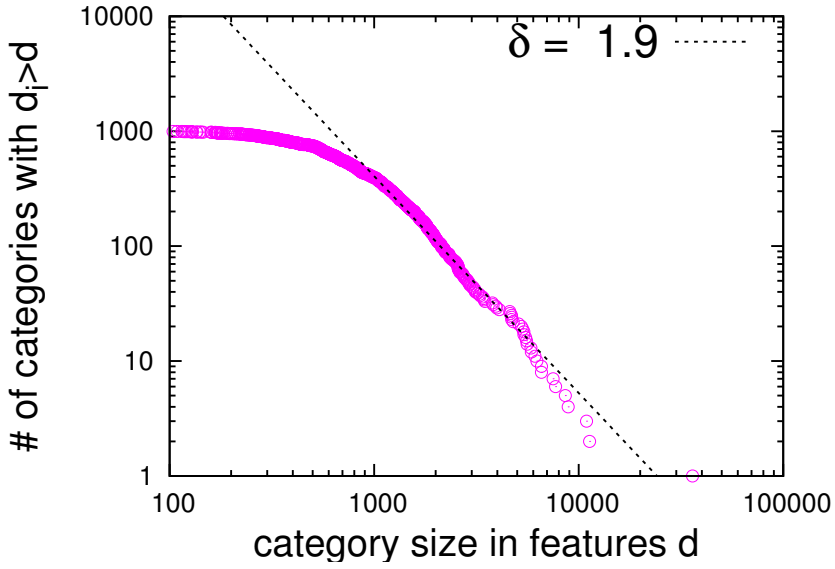

Figure 11: Number of features vs rank distribution.

classification and then we proceed to the space complexity analysis for large-scale systems. Finally, we empirically validate the derived bounds.

5.1 Hierarchical Classification

In single-label multi-class hierarchical classification, the training set can be represented by $S = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$. In the context of text classification, $\mathbf{x}^{(i)} \in \mathcal{X}$ denotes the vector representation of document i in an input space $\mathcal{X} \subseteq \mathbb{R}^d$.

The hierarchy in the form of rooted tree is given by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V} \supseteq \mathcal{Y}$ denotes the set of nodes of $\mathcal{G}$, and $\mathcal{E}$ denotes the set of edges with parent-to-child orientation. The leaves of the tree which usually form the set of target classes is given by $\mathcal{Y} = \{u \in \mathcal{V} : \nexists v \in \mathcal{V}, (u, v) \in \mathcal{E}\}$. Assuming that there are K classes, the label $y^{(i)} \in \mathcal{Y}$ represents the class associated with the instance $\mathbf{x}^{(i)}$. The hierarchical relationship among categories implies a transition from generalization to specialization as one traverses any path from root towards the leaves. This implies that the documents which are assigned to a particular leaf also belong to the inner nodes on the path from the root to that leaf node.

5.2 Space Complexity

The prediction speed for large-scale classification is crucial for its application in many scenarios of practical importance. It has been shown in [32, 3] that hierarchical classifiers are usually faster to train and test time as compared to flat classifiers. However, given the large physical memory of modern systems,

what also matters in practice is the size of the trained model with respect to the available physical memory. We, therefore, compare the space complexity of hierarchical and flat methods which governs the size of the trained model in large scale classification. The goal of this analysis is to determine the conditions under which the size of the hierarchically trained linear model is lower than that of flat model.

As a prototypical classifier, we use a linear classifier of the form $\mathbf{w}^T\mathbf{x}$ which can be obtained using standard algorithms such as Support Vector Machine or Logistic Regression. In this work, we apply one-vs-all $L2$-regularized $L2$-loss support vector classification as it has been shown to yield state-of-the-art performance in the context of large scale text classification [12]. For flat classification one stores weight vectors $\mathbf{w}_y, \forall y$ and hence in a K class problem in d dimensional feature space, the space complexity for flat classification is:

$$Size_{Flat} = d \times K \tag{12}$$

which represents the size of the matrix consisting of K weight vectors, one for each class, spanning the entire input space.

We need a more sophisticated analysis for computing the space complexity for hierarchical classification. In this case, even though the total number of weight vectors is much more since these are computed for all the nodes in the tree and not only for the leaves as in flat classification. Inspite of this, the size of hierarchical model can be much smaller as compared to flat model in the large scale classification. Intuitively, when the feature set size is high (top levels in the hierarchy), the number of classes is less, and on the contrary, when the number of classes is high (at the bottom), the feature set size is low.

In order to analytically compare the relative sizes of hierarchical and flat models in the context of large scale classification, we assume power law behaviour with respect to the number of features, across levels in the hierarchy. More precisely, if the categories at a level in the hierarchy are ordered with respect to the number of features, we observe a power law behaviour. This has also been verified empirically as illustrated in Figure 12 for various levels in the hierarchy, for one of the datasets used in our experiments. More formally, the feature size $d_{l,r}$ of the r-th ranked category, according to the number of features, for level l, $1 \leq l \leq L-1$, is given by:

$$d_{l,r} \approx d_{l,1} r^{-\beta_l} \tag{13}$$

where $d_{l,1}$ represents the feature size of the category ranked 1 at level l and $\beta > 0$ is the parameter of the power law. Using this ranking as above, let $b_{l,r}$ represent the number of children of the r-th ranked category at level l ($b_{l,r}$ is the branching factor for this category), and let B_l represents the total number of categories at level l. Then the size of the entire hierarchical classification model is given by:

$$Size_{Hier} = \sum_{l=1}^{L-1} \sum_{r=1}^{B_l} b_{l,r} d_{l,r} \approx \sum_{l=1}^{L-1} \sum_{r=1}^{B_l} b_{l,r} d_{l,1} r^{-\beta_l} \tag{14}$$

Here level $l = 1$ corresponds to the root node, with $B_1 = 1$.

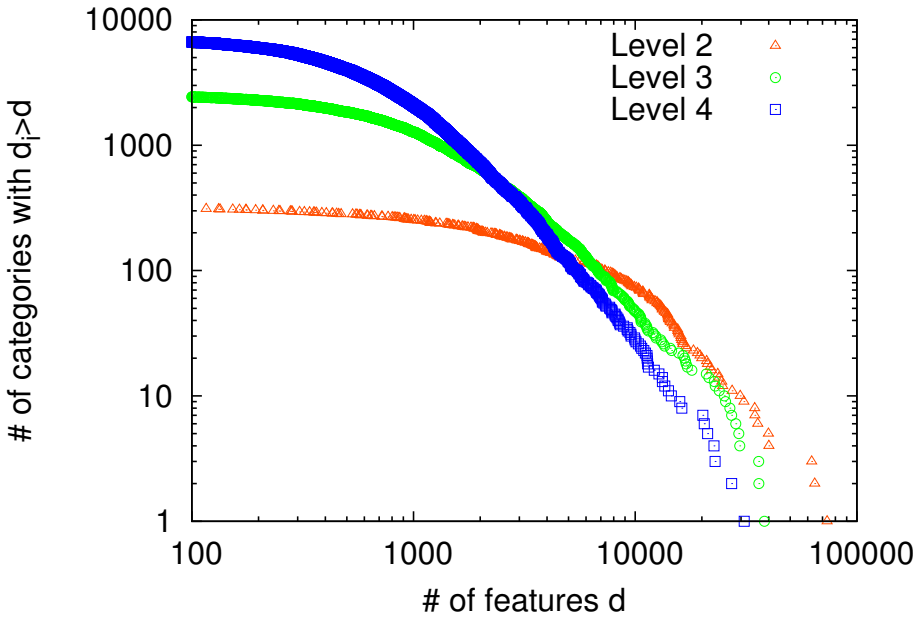

Figure 12: Power-law variation for features in different levels for LSHTC2-a dataset, Y-axis represents the feature set size plotted against rank of the categories on X-axis

We now state a proposition that shows that, under some conditions on the depth of the hierarchy, its number of leaves, its branching factors and power law parameters, the size of a hierarchical classifier is below that of its flat version.

Proposition 1 *For a hierarchy of categories of depth L and K leaves, let $\beta = \min_{1 \leq l \leq L} \beta_l$ and $b = \max_{l,r} b_{l,r}$. Denoting the space complexity of a hierarchical classification model by $Size_{hier}$ and the one of its corresponding flat version by $Size_{flat}$, one has:*

$$\text{For } \beta > 1, \text{if } \beta > \frac{K}{K - b(L-1)} (> 1), \text{ then } Size_{hier} < Size_{flat} \tag{15}$$

$$\text{For } 0 < \beta < 1, \text{if } \frac{b^{(L-1)(1-\beta)} - 1}{(b^{(1-\beta)} - 1)} < \frac{1-\beta}{b} K, \text{then } Size_{hier} < Size_{flat} \tag{16}$$

As $d_{l,1} \leq d_1$ and $B_l \leq b^{(l-1)}$ for $1 \leq l \leq L$, one has, from Equation 14 and the

definitions of β and b:

$$Size_{hier} \leq bd_1 \sum_{l=1}^{L-1} \sum_{r=1}^{b^{(l-1)}} r^{-\beta}$$

One can then bound $\sum_{r=1}^{b^{(l-1)}} r^{-\beta}$ using ([32]):

$$\sum_{r=1}^{b^{(l-1)}} r^{-\beta} < \left[\frac{b^{(l-1)(1-\beta)} - \beta}{1-\beta} \right] \text{ for } \beta \neq 0, 1 \tag{17}$$

leading to, for $\beta \neq 0, 1$:

$$\begin{aligned} Size_{hier} &< bd_1 \sum_{l=1}^{L-1} \left[\frac{b^{(l-1)(1-\beta)} - \beta}{1-\beta} \right] \\ &= bd_1 \left[\frac{b^{(L-1)(1-\beta)} - 1}{(b^{(1-\beta)} - 1)(1-\beta)} - (L-1)\frac{\beta}{(1-\beta)} \right] \end{aligned} \tag{18}$$

where the last equality is based on the sum of the first terms of the geometric series $(b^{(1-\beta)})^l$.

If $\beta > 1$, since $b > 1$, it implies that $\frac{b^{(L-1)(1-\beta)}-1}{(b^{(1-\beta)}-1)(1-\beta)} < 0$. Therefore, Inequality 18 can be re-written as:

$$Size_{hier} < bd_1(L-1)\frac{\beta}{(\beta-1)}$$

Using our notation, the size of the corresponding flat classifier is: $Size_{flat} = Kd_1$, where K denotes the number of leaves. Thus:

$$\text{If } \beta > \frac{K}{K - b(L-1)} (> 1), \text{then } Size_{hier} < Size_{flat}$$

which proves Condition 15.

The proof for Condition 16 is similar: assuming $0 < \beta < 1$, it is this time the second term in Equation 18 ($-(L-1)\frac{\beta}{(1-\beta)}$) which is negative, so that one obtains:

$$Size_{hier} < bd_1 \left[\frac{b^{(L-1)(1-\beta)} - 1}{(b^{(1-\beta)} - 1)(1-\beta)} \right]$$

and then:

$$\text{If } \frac{b^{(L-1)(1-\beta)} - 1}{(b^{(1-\beta)} - 1)} < \frac{1-\beta}{b} K, \text{then } Size_{hier} < Size_{flat}$$

which concludes the proof of the proposition.

It can be shown, but this is beyond the scope of this paper, that Condition 16 is satisfied for a range of values of $\beta \in]0, 1[$. However, as is shown in the experimental part, it is Condition 15 of Proposition 1 that holds in practice.

The previous proposition complements the analysis presented in [32] in which it is shown that the training and test time of hierarchical classifiers is importantly decreased with respect to the ones of their flat counterpart. In this work we show that the space complexity of hierarchical classifiers is also better, under a condition that holds in practice, than the one of their flat counterparts. Therefore, for large scale taxonomies whose feature size distribution exhibit power law decay, hierarchical classifiers should be better in terms of speed than flat ones, due to the following reasons:

1. As shown above, the space complexity of hierarchical classifier is lower than flat classifiers.

2. For K classes, only $O(\log K)$ classifiers need to be evaluated per test document as against $O(K)$ classifiers in flat classification.

In order to empirically validate the claim of Proposition 1, we measured the trained model sizes of a standard top-down hierarchical scheme (TD), which uses a linear classifier at each parent of the hierarchy, and the flat one.

We use the publicly available DMOZ data of the LSHTC challenge which is a subset of Directory Mozilla. More specifically, we used the large dataset of the LSHTC-2010 edition and two datasets were extracted from the LSHTC-2011 edition. These are referred to as LSHTC1-large, LSHTC2-a and LSHTC2-b respectively in Table 2. The fourth dataset (IPC) comes from the patent collection released by World Intellectual Property Organization. The datasets are in the LibSVM format, which have been preprocessed by stemming and stopword removal. Various properties of interest for the datasets are shown in Table 2.

Dataset	#Tr./#Test	#Classes	#Feat.
LSHTC1-large	93,805/34,880	12,294	347,255
LSHTC2-a	25,310/6,441	1,789	145,859
LSHTC2-b	36,834/9,605	3,672	145,354
IPC	46,324/28,926	451	1,123,497

Table 2: Datasets for hierarchical classification with the properties: Number of training/test examples, target classes and size of the feature space. The depth of the hierarchy tree for LSHTC datasets is 6 and for the IPC dataset is 4.

Table 3 shows the difference in trained model size (actual value of the model size on the hard drive) between the two classification schemes for the four datasets, along with the values defined in Proposition 1. The symbol $\bigtriangledown$ refers to the quantity $\frac{K}{K-b(L-1)}$ of condition 15.

As shown for the three DMOZ datasets, the trained model for flat classifiers can be an order of magnitude larger than for hierarchical classification. This results from the sparse and high-dimensional nature of the problem which is quite typical in text classification. For flat classifiers, the entire feature set

Dataset	TD	Flat	β	b	$\bigtriangledown$
LSHTC1-large	2.8	90.0	1.62	344	1.12
LSHTC2-a	0.46	5.4	1.35	55	1.14
LSHTC2-b	1.1	11.9	1.53	77	1.09
IPC	3.6	10.5	2.03	34	1.17

Table 3: Model size (in GB) for flat and hierarchical models along with the corresponding values defined in Proposition 1. The symbol $\bigtriangledown$ refers to the quantity $\frac{K}{K-b(L-1)}$

participates for all the classes, but for top-down classification, the number of classes and features participating in classifier training are inversely related, when traversing the tree from the root towards the leaves. As shown in Proposition 1, the power law exponent β plays a crucial role in reducing the model size of hierarchical classifier.

6 Conclusions

In this work we presented a model in order to explain the dynamics that exist in the creation and evolution of large-scale taxonomies such as the DMOZ directory, where the categories are organized in a hierarchical form. More specifically, the presented process models jointly the growth in the size of the categories (in terms of documents) as well as the growth of the taxonomy in terms of categories, which to our knowledge have not been addressed in a joint framework. From one of them, the power law in category size distribution, we derived power laws at each level of the hierarchy, and with the help of Heaps's law a third scaling law in the features size distribution of categories which we then exploit for performing an analysis of the space complexity of linear classifiers in large-scale taxonomies. We provided a grounded analysis of the space complexity for hierarchical and flat classifiers and proved that the complexity of the former is always lower than that of the latter. The analysis has been empirically validated in several large-scale datasets showing that the size of the hierarchical models can be significantly smaller that the ones created by a flat classifier.

The space complexity analysis can be used in order to estimate beforehand the size of trained models for large-scale data. This is of importance in large-scale systems where the size of the trained models may impact the inference time.

7 Acknowledgements

This work has been partially supported by ANR project Class-Y (ANR-10-BLAN-0211), BioASQ European project (grant agreement no. 318652), LabEx

PERSYVAL-Lab ANR-11-LABX-0025, and the Mastodons project Garguantua.

References

[1] R. Babbar, I. Partalas, C. Metzig, E. Gaussier, and M.-R. Amini. Comparative classifier evaluation for web-scale taxonomies using power law. In *European Semantic Web Conference*, 2013.

[2] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[3] S. Bengio, J. Weston, and D. Grangier. Label embedding trees for large multi-class tasks. In *Neural Information Processing Systems*, pages 163–171, 2010.

[4] P. N. Bennett and N. Nguyen. Refined experts: improving classification in large taxonomies. In *Proceedings of the 32nd international ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 11–18, 2009.

[5] L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In *Advances In Neural Information Processing Systems*, pages 161–168, 2008.

[6] L. Cai and T. Hofmann. Hierarchical document categorization with support vector machines. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 78–87, 2004.

[7] A. Capocci, V. D. Servedio, F. Colaiori, L. S. Buriol, D. Donato, S. Leonardi, and G. Caldarelli. Preferential attachment in the growth of social networks: The internet encyclopedia wikipedia. *Physical Review E*, 74(3):036116, 2006.

[8] O. Dekel, J. Keshet, and Y. Singer. Large margin hierarchical classification. In *Proceedings of the twenty-first international conference on Machine learning*, ICML '04, pages 27–34, 2004.

[9] S. N. Dorogovtsev and J. F. F. Mendes. Evolution of networks with aging of sites. *Physical Review E*, 62(2):1842, 2000.

[10] L. Egghe. Untangling herdan's law and heaps' law: Mathematical and informetric arguments. *Journal of the American Society for Information Science and Technology*, 58(5):702–709, 2007.

[11] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On power-law relationships of the internet topology. *SIGCOMM*.

[12] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[13] T. Gao and D. Koller. Discriminative learning of relaxed hierarchy for large-scale visual recognition. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2072–2079, 2011.

[14] M. M. Geipel, C. J. Tessone, and F. Schweitzer. A complementary view on the growth of directory trees. *The European Physical Journal B*, 71(4):641–648, 2009.

[15] S. Gopal, Y. Yang, B. Bai, and A. Niculescu-Mizil. Bayesian models for large-scale hierarchical classification. In *Neural Information Processing Systems*, 2012.

[16] G. Jona-Lasinio. Renormalization group and probability theory. *Physics Reports*, 352(4):439–458, 2001.

[17] K. Klemm, V. M. Eguíluz, and M. San Miguel. Scaling in the structure of directory trees in a computer cluster. *Physical review letters*, 95(12):128701, 2005.

[18] D. Koller and M. Sahami. Hierarchically classifying documents using very few words. In *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, 1997.

[19] T.-Y. Liu, Y. Yang, H. Wan, H.-J. Zeng, Z. Chen, and W.-Y. Ma. Support vector machines classification with a very large-scale taxonomy. *SIGKDD*, 2005.

[20] B. Mandelbrot. A note on a class of skew distribution functions: Analysis and critique of a paper by ha simon. *Information and Control*, 2(1):90–99, 1959.

[21] C. Metzig and M. B. Gordon. A model for scaling in firms' size and growth rate distribution. *Physica A*, 2014.

[22] M. Newman. Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46(5):323–351, 2005.

[23] M. E. J. Newman. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 2005.

[24] I. Partalas, R. Babbar, É. Gaussier, and C. Amblard. Adaptive classifier selection in large-scale hierarchical classification. In *ICONIP*, pages 612–619, 2012.

[25] P. Richmond and S. Solomon. Power laws are disguised boltzmann laws. *International Journal of Modern Physics C*, 12(03):333–343, 2001.

[26] H. A. Simon. On a class of skew distribution functions. *Biometrika*, 42(3/4):425–440, 1955.

[27] C. Song, S. Havlin, and H. A. Makse. Self-similarity of complex networks. *Nature*, 433(7024):392–395, 2005.

[28] H. Takayasu, A.-H. Sato, and M. Takayasu. Stable infinite variance fluctuations in randomly amplified langevin systems. *Physical Review Letters*, 79(6):966–969, 1997.

[29] C. J. Tessone, M. M. Geipel, and F. Schweitzer. Sustainable growth in complex networks. *EPL (Europhysics Letters)*, 96(5):58005, 2011.

[30] K. G. Wilson and J. Kogut. The renormalization group and the expansion. *Physics Reports*, 12(2):75–199, 1974.

[31] G.-R. Xue, D. Xing, Q. Yang, and Y. Yu. Deep classification in large-scale text hierarchies. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 619–626, 2008.

[32] Y. Yang, J. Zhang, and B. Kisiel. A scalability analysis of classifiers in text categorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 96–103, 2003.

[33] G. U. Yule. A mathematical theory of evolution, based on the conclusions of dr. jc willis, frs. *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character*, 213:21–87, 1925.